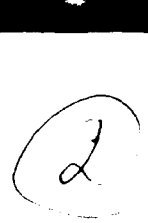


Segment-based Acoustic Models for Continuous Speech Recognition

Progress Report: 1 October 92 – 30 September 93

submitted to
Office of Naval Research
and
Advanced Research Projects Administration
8 October 1993

by
Boston University
Boston, Massachusetts 02215

Principal Investigators

Dr. Mari Ostendorf
Assistant Professor of ECS Engineering, Boston University
Telephone: (617) 353-5430

Dr. J. Robin Rohlicek
Scientist, BBN Inc.
Telephone: (617) 873-3894

Administrative Contact

Maureen Rogers, Awards Manager
Office of Sponsored Programs
Telephone: (617) 353-4365

Executive Summary

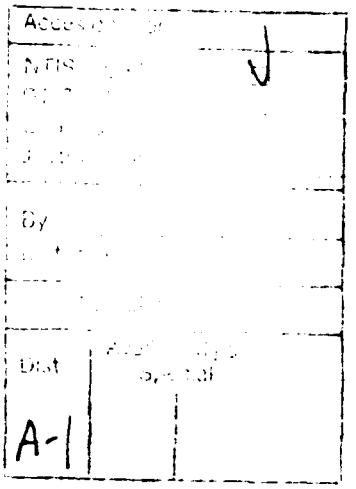
This research aims to develop new and more accurate stochastic models for speaker-independent continuous speech recognition by extending previous work in segment-based modeling and by introducing a new hierarchical approach to representing intra-utterance statistical dependencies. These techniques, which have high computational costs because of the large search space associated with higher order models, are made feasible through rescoring a set of HMM-generated N-best sentence hypotheses. We expect these different modeling techniques to result in improved recognition performance over that achieved by current systems, which handle only frame-based observations and assume that these observations are independent given an underlying state sequence.

In the past year, the accomplishments of this project, funded in part by a related ARPA-NSF grant (NSF no. IRI-8902124), include:

- improved the N-best rescoring paradigm by introducing score normalization and more robust weight estimation techniques;
- investigated techniques for improving the baseline stochastic segment model (SSM) system, including context clustering for robust parameter estimation, tied mixture distributions at the frame and segment level, a two level segment/microsegment formalism, multiple pronunciation word models, and automatic distribution mapping estimation;
- extended the classification and segmentation scoring formalism to context-dependent modeling without assuming independence of observations in different segments, which opens the possibility for a broader class of features for recognition;
- demonstrated results comparable to the best HMM systems on the Resource Management, Switchboard and Wall Street Journal tasks;
- developed an initial dependency tree model of intra-utterance observation correlation; and
- implemented and evaluated baseline n-gram language models, and developed new language models to handle topic-related language dynamics and variations in verbalized numbers and punctuation.

We currently report baseline SSM results on the Wall Street Journal task that represent improved performance over all results reported in November 1992. For the 5k vocabulary, non-verbalized punctuation test set and the bigram language model, we achieve 8.1% error with the SSM and 7.3% error with the combined HMM-SSM system, which can be compared to reported rates of 8.7% - 15% for comparable HMM systems. In addition, we see much room for further improvement, as these models still rely on an assumption of conditional independence and do not take full advantage of the segment formalism.

Contents



Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 Oct 1992 - 30 September 1993

1 Productivity Measures

- Refereed papers submitted but not yet published: 1
- Refereed papers published: 1
- Unrefereed reports and articles: 3
- Books or parts thereof submitted but not yet published: 0
- Books or parts thereof published: 0
- Patents filed but not yet granted: 0
- Patents granted (include software copyrights): 0
- Invited presentations: 0
- Contributed presentations: 1 talk, 1 poster
- Honors received:
  Prof. M. Ostendorf: Served on the IEEE Signal Processing Society Speech Technical Committee; Chosen to chair the 1996 ARPA Workshop on Human Language Technology; Invited to participate in the DoD workshop, Frontiers in Speech Processing - Robust Speech Recognition. Dr. J. R. Rohlicek: Chosen to serve as an Associate Editor for *IEEE Signal Processing Letters*.
- Prizes or awards received: 0
- Promotions obtained: At Boston University, Prof. Ostendorf was granted tenure and promoted from Assistant Professor to Associate Professor. At BBN, Dr. Rohlicek was promoted to Division Scientist.
- Graduate students supported $\geq$ 25% of full time: 2-4
- Post-docs supported $\geq$ 25% of full time: 0
- Minorities supported: 1 woman

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 April 1993 - 30 June 1993

2 Summary of Technical Progress

In this work, we are interested in the problem of large vocabulary, speaker-independent continuous speech recognition, and primarily in the acoustic modeling component of this problem. In developing acoustic models for speech recognition, we have conflicting goals. On one hand, the models should be robust to inter- and intra-speaker variability, to the use of a different vocabulary in recognition than in training, and to the effects of moderately noisy environments. In order to accomplish this, we need to model gross features and global trends. On the other hand, the models must be sensitive and detailed enough to detect fine acoustic differences between similar words in a large vocabulary task. To answer these opposing demands requires improvements in acoustic modeling at several levels: the frame level (e.g. signal processing), the phoneme level (e.g. modeling feature dynamics), and the utterance level (e.g. defining a structural context for representing the intra-utterance dependence across phonemes). This project addresses the problem of acoustic modeling, specifically focusing on modeling at the segment level and above. The research strategy includes three main thrusts. First, phone-level acoustic modeling is based on the stochastic segment model (SSM), and in this area our main efforts involve developing new techniques for robust context modeling, mechanisms for effectively incorporating segmental features, and models of within-segment dependence of frame-based features. Second, high-level models are being explored in order to capture speaker-dependent and session-dependent effects within the context of a speaker-independent model. In particular, we are investigating hierarchical structures for representing the intra-utterance dependency of phonetic models, and more recently language models for representing topic dependency and language dynamics, recognizing that higher-order models of correlation can extend to the language domain as well as the acoustic domain. Lastly, speech recognition is implemented under the N-best rescoring paradigm, in which the BBN Byblos system is used to constrain the stochastic segment model (SSM) search space by providing the top N sentence hypotheses. This paradigm facilitates research on high-order models through reducing development costs, and provides a modular framework for technology transfer that has already enabled us to advance state-of-the-art recognition performance through collaboration with BBN.

In the first year of this project, we have focused on improving the performance of the basic segment word recognition system and porting the system to the Wall Street Journal task domain.

The different accomplishments and advances, some of which were supported in part by an ARPA-NSF grant (NSF no. IRI-8902124), are detailed below.

N-Best rescoring. We developed a grid-based search to avoid local optima in the weight optimization criterion, together with methods for choosing among different local optima to obtain more robust results. We also found that normalization of scores by observation length (e.g., frame, phoneme, or word count, depending on the score) prior to the linear combination allows us to obtain more robust weights and leads to a small reduction in error rate.

Improvements to the SSM. We focused on improving the performance of the basic segment word recognition system. In brief, the accomplishments of that period include the following: 1) development of a method for clustering contexts to provide robust context-dependent model parameter estimates using a likelihood ratio test to obtain ML estimates of tied covariances, obtaining a factor of 10 reduction in memory costs with no loss in performance; 2) extension of the two level segment/microsegment formalism (to context-dependent modeling in word recognition) and assessment of trade-offs in mixture vs. trajectory modeling, finding that (non-tied) mixtures are more useful for context-independent modeling and constrained trajectories are more appropriate for context-dependent modeling; 3) investigation of the use of tied mixtures at the frame level and at the segment level, looking at trade-offs of different methods for parameter initialization and different regions of parameter tying, achieving a 20% reduction in word error on the RM task by using frame-level full covariance tied mixtures, but no gains on the WSJ task; 4) development and assessment of automatically generated multiple-pronunciation word networks (no performance improvements obtained in experiments on the Resource Mangagement task, but higher quality phone alignments are obtained in other tasks); 5) implementation of optional silence insertion in both recognition and training, which led to a slight improvement in performance on WSJ; and 6) automatic distribution mapping estimation using a maximum likelihood criterion, which is an important development needed for extending the segment model to different speech units and different feature sets.

CIR framework. One approach to segment-based modeling is to do "classification in recognition" (CIR), or classification of a variable-length segment using a posterior distribution based on fixed-length features, a useful formalism because it opens the possibility for a broader class of features for recognition. In the past, we have shown that this approach requires both classification and segmentation scoring to be effective. In this project, we made an important step forward in building a formalism for using posterior distributions in classification through our development of a mechanism to handle context-dependent models without requiring the assumption of independence of features spanning different phone segments. The context-dependent model was derived using a maximum entropy criterion in estimating a combined function of posterior probability terms. This formalism will allow the use of acoustic measurements over a longer time span and facilitate hierarchical modeling. Through mathematical analysis as well as experiments in context-dependent modeling, we uncovered fundamental problems in reported implementations of context-dependent

CIR scoring, that require changes to the classification score.

SSM baseline results. We have ported the baseline SSM to the Resource Management, Switchboard and Wall Street Journal tasks, and demonstrated speaker independent recognition results comparable to the best HMM systems. On the *Resource Management* (RM) task we report 3.6% word error on the October 1989 Resource Management test for the SSM alone, and 3.1% word error for the combined SSM-HMM system. (The best reported HMM result on this test set is LIMSI's 3.2% error rate.) On the September 1992 test set for this task, our performance figures are 7.3% and 6.1% word error for these two systems, which are also very good results given the difficulty of the test set. We ported our recognition system to the *Switchboard* credit card task, as part of our participation in the Robust Speech Recognition Workshop at Rutger's this past summer. Our results of 29% accuracy for gender-independent models was comparable to all HMM systems reporting on this task, excluding the 32-33% accuracy achieved by systems using gender-dependent models. Our baseline results on the *Wall Street Journal* (WSJ) 5k vocabulary task represent improved performance over all results reported in November 1992. For the non-verbalized punctuation test set and the bigram language model, we achieve 8.1% error with the SSM and 7.3% error with the combined HMM-SSM system, which can be compared to reported rates of 8.7% - 15% for comparable HMM systems. Interestingly, our best results on the WSJ task are based on full-covariance, single-mode Gaussians, while the best results on the RM task are achieved with tied-mixture models. The RM results of the context-clustering algorithm were confirmed on the WSJ task.

Dependency tree model. An important goal of this project is the development of a hierarchical model of intra-utterance correlation of phone observations. Our initial efforts in this area have been to extend prior work on finding the minimal spanning dependence trees, from the discrete models of Chow and Liu to Gauss-Markov models of dependence. The initial implementation favored connections between infrequently observed classes, so we are currently investigating robust algorithms for designing trees, as well as the use of discrete distribution dependence trees in mixture models. In order to quickly assess different models of dependence without the high cost of building a full word recognition system, we are initially comparing prediction errors for different models within the context of the TIMIT corpus.

Language models. Motivated by the realization that inter- and intra-utterance correlation can be modeled at the language as well as acoustic level, we have begun an effort in dynamic language modeling. As a first step in this project, we have implemented the Katz and Witten-Bell back-off algorithms for estimating n-gram language models, and are currently evaluating the impact of these differences on recognition performance. We developed a formalism for modeling the probability of the different alternatives people have for verbalizing numbers and punctuation, recognizing that in spontaneous dictation, some types of punctuation are more likely to be verbalized than others. Finally, we developed a mixture language model formalism that represents the topic-dependent structure of language at the utterance level.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 April 1993 - 30 June 1993

3 Publications and Presentations

Papers associated with this work and written during the reporting period include a site report, two conference papers, and a correspondence paper that was submitted and accepted for publication during the reporting period. A journal paper documenting our prior work in recognition also appeared during this period.

- "Fast Search Algorithms for Phone Classification and Recognition Using Segment-Based Models," V. Digalakis, M. Ostendorf and J. R. Rohlicek, *IEEE Transactions on Signal Processing*, December 1992, pp. 2885-2896.
- "Segment-Based Acoustic Models for Continuous Speech Recognition," M. Ostendorf and J. R. Rohlicek, site report to appear in *Proceedings of the ARPA Workshop on Human Language Technology*, 1993.
- "On the Use of Tied-Mixture Distributions," O. Kimball and M. Ostendorf, to appear in *Proceedings of the ARPA Workshop on Human Language Technology*, 1993.
- "A Comparison of Trajectory and Mixture Modeling in Segment-based Word Recognition," A. Kannan and M. Ostendorf, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. II327-330, April 1993.
- "Maximum Likelihood Clustering of Gaussians for Speech Recognition," A. Kannan, M. Ostendorf and J. R. Rohlicek, *IEEE Transactions on Speech and Audio Processing*, to appear.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 April 1993 - 30 June 1993

4 Transitions and DoD Interactions

This grant includes a subcontract to BBN, and the research results and software is available to them. Thus far, we have collaborated with BBN by combining the Byblos system with the SSM in N-Best sentence rescoring to obtain improved recognition performance, and we have provided BBN with papers and technical reports to facilitate sharing of algorithmic improvements. On their part, BBN has been very helpful to us in our WSJ porting efforts, providing us with WSJ data and consulting on format changes.

The recognition system that has been developed under the support of this grant and of a joint NSF-ARPA grant (NSF # IRI-8902124) is currently being used for automatically obtaining good quality phonetic alignments for a corpus of radio news speech under development at Boston University. The alignment effort is supported by the Linguistic Data Consortium, through a grant that allowed us to add cross-word phonological rules to the segmentation software.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 April 1993 – 30 June 1993

5 Software and Hardware Prototypes

Our research has required the development and refinement of software systems for parameter estimation and recognition search, which are implemented in C or C++ and run on Sun Sparc workstations. No commercialization is planned at this time.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 April 1993 – 30 June 1993

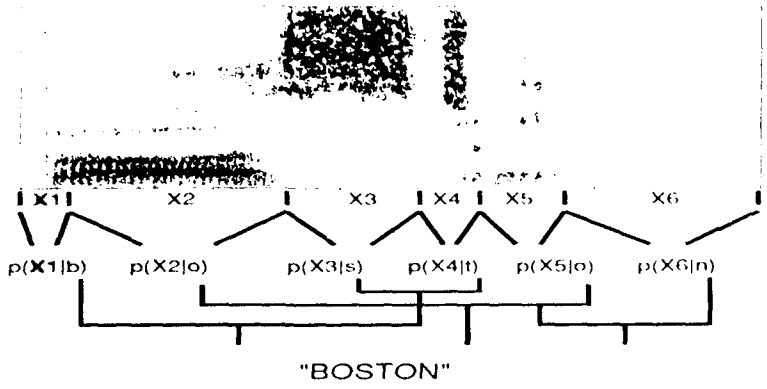
6 Vugraphs

Attached is a quad chart that illustrates the acoustic modeling philosophy of this project and lists the goals and key accomplishments. This chart was provided to ARPA earlier this year.

Segment-Based Acoustic Models for Continuous Speech Recognition

Mari Ostendorf, Boston University

J. Robin Rohlicek, BBN Inc.



New Ideas

- Use of parameter tying to capture context in non-traditional stochastic phone models
- Use of dependency trees to represent intra-utterance correlation of model parameters
- Evaluation of hierarchical models in adaptation and multi-pass recognition search.

Impact

Improvements in acoustic modeling are fundamental to speech processing, i.e.,

- speech recognition
- word spotting

and will impact many DOD applications, including

- database query
- command and control
- gisting

Schedule